# Bioinformatics for Medical Diagnostics: Assessment of Microarray Data in the Context of Clinical Databases

**Dugas M[1], Merk S[1], Breit S[2], Schoch C[3], Haferlach T[3], Kääb S[4]**

[1]Department of Medical Informatics, [2]Department of Dermatology, [3]Department of Internal Medicine III, [4]Department of Internal Medicine I, University of Munich, Germany

## ABSTRACT

*Motivation*: To identify genes suitable for medical diagnostics microarray data is assessed in the context of clinical databases, which store complex information about the patient phenotype. The wealth of data and lack of standards make it difficult to analyse this kind of data.

*Results*: We present a workflow for exploratory analysis of microarray data together with clinical data consisting of four steps: definition of clinically meaningful research questions in a masterfile, generation of analysis files, selection and characterization of differentially expressed genes, and estimation of classification accuracy. We applied this workflow to large data sets from the field of cardiology and oncology (n~500 patients).

Systematic data management of microarray data and clinical data helps to make results more transparent and comparable.

*Keywords*: data management, DNA microarray, medical diagnostics, clinical database, gene expression

## INTRODUCTION

Microarrays are being applied to investigate diseases on a molecular level. The interpretation of the data is difficult because the number of measurement points is much higher than the number of samples and the correlation structure of the gene expression levels is unknown. For medical diagnostics differentially expressed genes, particularly disease-specific genes, are a major focus of ongoing research [1,2].

To interprete microarray data from patient samples, integration with clinical data is required, e.g. follow-up information concerning patient survival. By integrating information from different diagnostic modalities (clinical classification, laboratory diagnostics, especially PCR) the medical plausibility and consistency of microarray data can be verified.

## SYSTEMS AND METHODS

Clinical databases are characterized by a large number of attributes (typically >100 per patient) and many different medical coding schemes [3,4]. For this reason there are many possibilities to partition the same set of patients into different clinically meaningful groups, e.g. young versus old patients, patients with high versus low blood pressure, patients with normal versus abnormal cholesterol level etc.

Typically, different cut-off values are possible to define a "high" or "low" value of an attribute, and usually there are not two (A versus B), but 5 - 10 medical categories (e.g. diagnoses).

Clinically relevant groups are usually defined by a medical expert. They can be determined by a single attribute, e.g. location of the disease, or by a combination of several attributes, e.g. high blood pressure combined with high blood glucose as a high risk group of patients.

As a consequence, many different clinically meaningful research questions can be posed for the same data set. For each research question, a list of differentially expressed genes, gene profiles with annotation and an estimation of diagnostic accuracy need to be determined.

Our approach to assess microarray data in the context of clinical databases consists of the following steps:

### 1. Definition of clinically meaningful research questions in a masterfile

This definition is performed by a medical expert. He selects attributes from the clinical database, which are important to assign the patients to clinically meaningful groups. This assignment is based on single attributes or a combination of attributes.

Table 1 shows the structure of the masterfile.

The medical rationale, the rules and the cut-off values to define categories are subject to medical expertise and are documented separately. In our setting, up to 30 analysis columns (=research questions) per data set and up to 25 categories per analysis column were defined.

| sampleID | attribute-1 | attribute-2 | ... | attribute-n | analysis -1 | analysis -2 | ... | analysis -n |
|---|---|---|---|---|---|---|---|---|
| microarray1 | 10 | high | | 20 | normal | location1 | | intermediate |
| microarray2 | 120 | high | | 1000 | disease1 | location1 | | severe |
| microarray3 | 120 | high | | 1230 | disease1 | location1 | | severe |
| microarray4 | 500 | low | | 1120 | disease2 | location2 | | intermediate |
| microarray5 | 500 | low | | 30 | disease2 | location2 | | mild |
| microarray6 | 5 | low | | 23 | normal | location2 | | mild |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |

Table 1: Masterfile. Each row corresponds to a patient sample, identified by a sample identifier (sampleID). Columns attribute-1 to attribute-n denote selected clinical attributes from the patient database (like blood pressure, smoking status etc.). Single attributes or certain combinations of attributes define patient groups for microarray analysis. For each research question, an analysis column (analysis -1 to analysis -n) is provided by the medical expert.

| geneID | normal | disease1 | disease1 | disease2 | disease2 | normal | ... |
|---|---|---|---|---|---|---|---|
| | microarray1 | microarray2 | microarray3 | microarray4 | microarray5 | microarray6 | ... |
| gene-1 | 4 | 132 | 17 | 55 | 66 | 44 | ... |
| gene-2 | 46 | 23 | 2 | 2344 | 44 | 55 | ... |
| ... | ... | ... | ... | ... | ... | ... | ... |
| gene-n | 86 | 43 | 54 | 55 | 75 | 34 | ... |

Table 2: Analysis file. The first column consists of a geneID. The first row provides the diagnostic categories of the samples, the second row lists sample identifiers. Starting from the third row, the gene expression values are provided.

## 2. Generation of analysis files

A straightforward PERL-Script (http://www.perl.com) generates a separate analysis file (Table 2) for each research question, defined by the masterfile. Raw data files and masterfiles are linked by means of sample identifiers. The analysis file is divided into a training set, which consists of two thirds of samples, and a test set, which contains the rest of the data.

## 3. Selection and characterization of differentially expressed genes

There are several published methods to identify differentially expressed genes in microarray data sets with various advantages and disadvantages. To unify the analysis process and to provide both sensitive and specific methods we apply three established methods to identify differentially expressed genes in the training sets: maxT-minP according to Westfall & Young, Golub's neighborhood analysis and False discovery rate (for details see [5]). We also apply the R-packages from Bioconductor (http://www.bioconductor.org, [6]). Each method generates a list of gene identifiers together with parameters such as adjusted p-values.

The three lists of differentially expressed genes provided by these methods are combined automatically and enhanced with annotation by means of PERL-programs. For each gene, information is provided concerning the statistical method to which it is assigned per differential expression. In addition, gene expression profiles are generated.

## 4. Estimation of classification accuracy

The merged gene list generated in step 3 and the gene expression data from each training set are used to train a support vector machine (SVM, [7]). The samples in each test set are classified using this SVM-model and the classification accuracy is estimated.

The workflow for this systematic analysis process of gene expression data, in conjunction with clinical databases, is summarized in Figure 1.

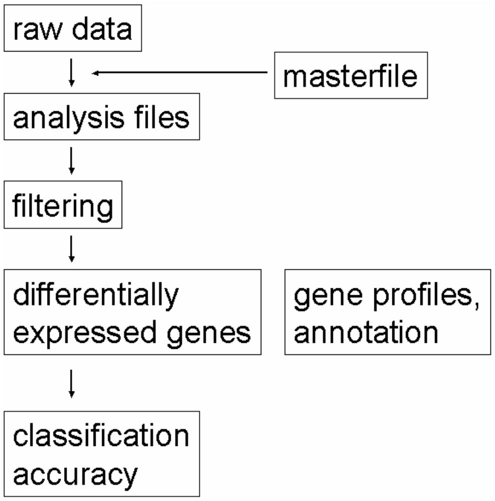
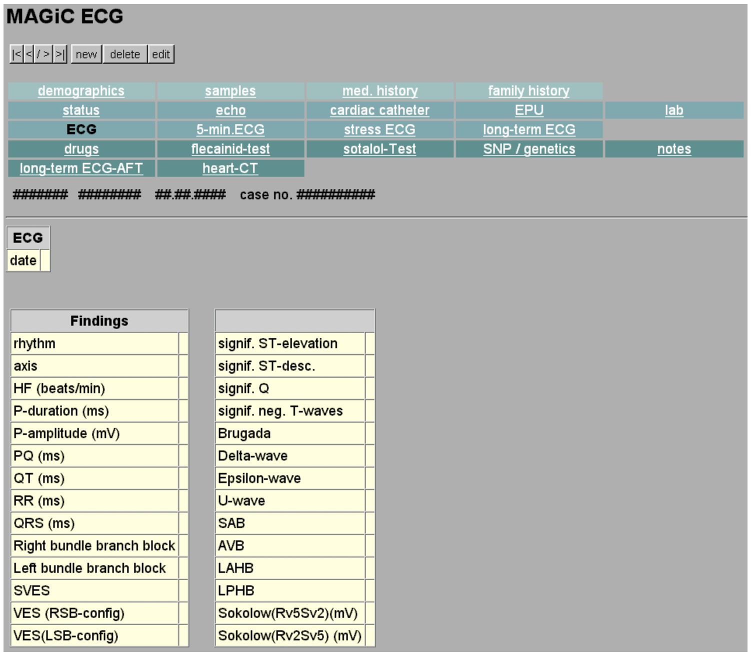


Figure 1: Workflow for systematic gene expression analysis.



####### ######## ##.##.#### case no. ##########

| ECG |
|---|
| date |

| Findings | | |
|---|---|---|
| rhythm | signif. ST-elevation | |
| axis | signif. ST-desc. | |
| HF (beats/min) | signif. Q | |
| P-duration (ms) | signif. neg. T-waves | |
| P-amplitude (mV) | Brugada | |
| PQ (ms) | Delta-wave | |
| QT (ms) | Epsilon-wave | |
| RR (ms) | U-wave | |
| QRS (ms) | SAB | |
| Right bundle branch block | AVB | |
| Left bundle branch block | LAHB | |
| SVES | LPHB | |
| VES (RSB-config) | Sokolow(Rv5Sv2)(mV) | |
| VES(LSB-config) | Sokolow(Rv2Sv5) (mV) | |

Figure 2: ECG-documentation.

## IMPLEMENTATION

We built complex clinical databases in the field of cardiology and oncology, consisting of 500–1000 attributes per patient and analysed gene expression by means of Affymetrix$^{TM}$ microarrays (U95a and U133a, n~500). Figure 2 presents a screenshot of electrocardiogram (ECG) data from this cardiological database (MAGiC=Munich Alliance for Genomic research on Cardiac arrhythmias). This system comprises of 20 documentation modules representing a detailed phenotype of cardiological patients.

Figure 3 presents two profiles of genes which were detected as being differentially expressed between patient groups. It becomes evident that a gene with significantly different means between groups is not necessarily able to separate groups precisely, i.e. a specific gene for a certain patient category.

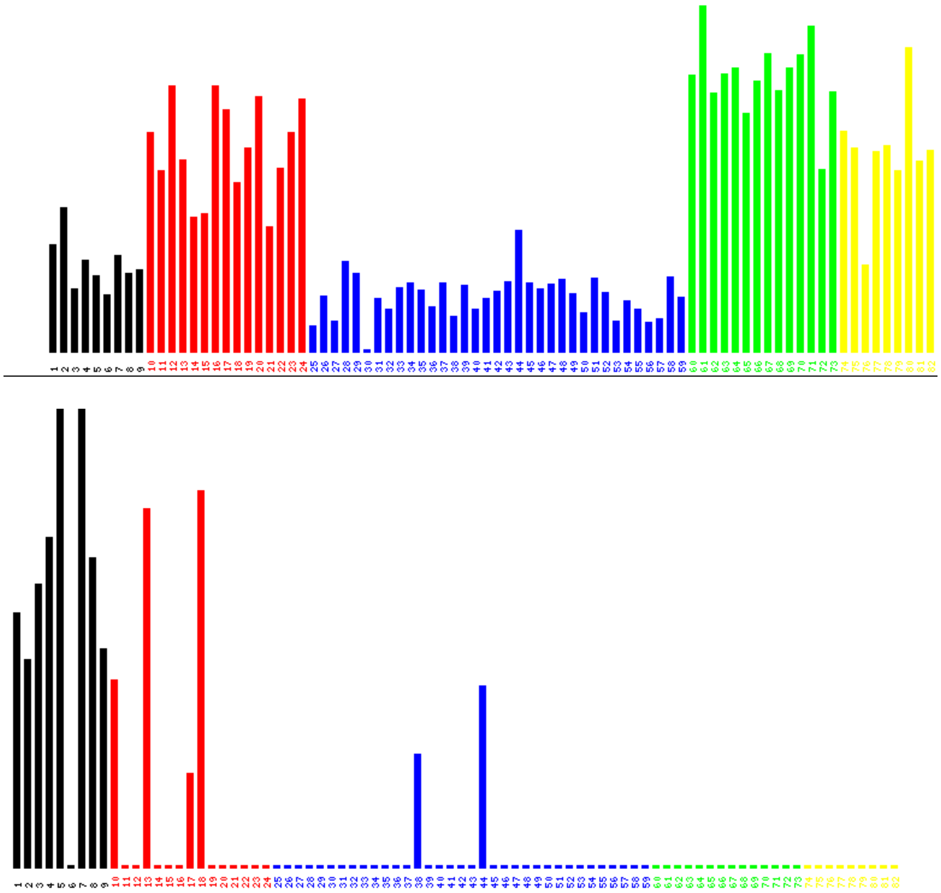


Figure 3: Gene profile for two different genes. Each bar corresponds to the gene expression level of one gene in one microarray experiment (height of the bars: arbitrary units). Each color corresponds to a certain patient category. The red and the blue group can be separated in the the upper, but not in the lower gene profile.

## DISCUSSION AND CONCLUSION

It is well known that microarrays generate a lot of data. To make it even worse, in a clinical setting – due to the complexity of the patient phenotype – many research questions can be posed of these data sets. Therefore, a systematic approach of data management is useful to keep an overview and to avoid errors in analysis.

We developed a method to manage clinical and microarray data in a real-world clinical setting with several hundred patients, dozens of research questions and up to 25 patient categories per analysis. The medical expert defined clinically meaningful research questions in a masterfile, which was used to generate analysis files. An analysis workflow suitable for large-scale microarray analysis was defined to answer medically important questions in a systematic manner: Which genes are differentially expressed? What is the estimated classification accuracy for a diagnostic test based on these genes? These questions are important because differential expression of a gene is not equivalent to disease-specificity, which is relevant for a diagnostic setting.

We applied established methods for data processing and analysis from the literature. In our case, the results were encouraging because many findings were consistent with genetic phenomena known from the literature and available RT-PCR data; in addition, the results matched well with different microarray types [8].

However, there are important problems. This kind of analysis is highly exploratory and worsens the multiple testing problem. Therefore, confirmatory experiments and validation are needed. There is an urgent need for standardisation of methods for detecting differentially expressed genes and for estimation of classification accuracy. This should address both selection of methods and guidelines for reasonable parameters. In particular, scientists are currently debating which method is most appropriate for differentially expressed genes. For this reason, we applied three established methods. From our experience, the main results from these methods are similar if the data set is "very clear". For estimation of classification accuracy we used SVM, because it usually outperforms other methods [9]. We did not address the problem of normalization and applied the default procedure from the microarray manufacturer.

Further research is needed to detect structures in this kind of data in a reliable manner. Systematic data management is important to streamline these analyses and to make the results more transparent and comparable. When these problems are solved, there is much potential for application of microarray technology in the field of medical diagnostics.

## ACKNOWLEDGEMENTS

Supported by grants from the German Ministry for Education and Research (BMBF): National Genome Research Network (NGFN) - Munich Alliance for Genomic Research on Cardiac Arrythmias (01GS0109) and Kompetenznetz Akute und Chronische Leukämien (01 GI 9980/6). Thanks to Dr. J. Kerr for grammar revision.

## REFERENCES

1. van de Vijver, M.J., Yudong, H., et al. (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999-2009

2. Schoch, C., Kohlmann, A., et al. (2002) Acute myeloid leukemias with reciprocal rearrangements can be distinguished by specific gene expression profiles. PNAS 99, 10008–10013

3. Dugas, M., Hoffmann, E., et al. (2002) Complexity of Biomedical Data Models in Cardiology: The Intranet-based Atrial Fibrillation Registry. Comput Methods Programs Biomed. 68, 49-61

4. Dugas, M., Schoch, C., et al. (2001) A comprehensive leukemia database: Integration of cytogenetics, molecular genetics and microarray data with clinical information, cytomorphology and immunphenotyping. Leukemia 15, 1805-1810

5. Dudoit, S., Fridlyand, J., Speed, T.P. (2002) Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. JASA 97:77-87

6. Gentleman, R. and Carey, V. (2002) Bioconductor. R News, 2(1), 11–16

7. Furey, T.S., Cristianini, N., et al. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 10:906-914

8. Kohlmann, A., Schoch, C., et al. (2003) Molecular characterization of acute leukemias by use of microarray technology. Genes, Chromosomes and Cancer 37:396-405

9. Yeang, C.H., Ramaswamy, S., Tamayo, P., et al. (2001) Molecular classification of multiple tumor types. Bioinformatics 17(Suppl1),S316-322